(2)

AD-A235 581

1. REPORT NUMBER: #56	READ INSTRUCTIONS BEFORE COMPLETING FORM 2. GOVT ACCESSION NO. 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): Feature Extraction Using an Unsupervised Neural Network	5. TYPE OF REPORT & PERIOD COVERED: Technical 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Nathan Intrator	8. CONTRACT OR GRANT NUMBER(s): N00014-86-K-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Brown University Center for Neural Science Providence, Rhode Island 02912	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: N-201-484
11. CONTROLLING OFFICE NAME AND ADDRESS: Personnel & Training Rsch. Program Office of Naval Research, Code 442PT Arlington, Virginia 22217	12. REPORT DATE: May 3, 1991 13. NUMBER OF PAGES: 10 pages
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report) 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

DTIC

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited. Publication in part or in whole is permitted for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Feature Extraction, Dimensionality reduction Neural Network, Classification, Speech recognition

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A novel unsupervised neural network for dimensionality reduction which seeks directions emphasizing distinguishing features in the data is presented. A satistical framework for the parameter estimation problem associated with this neural netowrk is given and its connection to exploratory projection pursuit methods is established.

Feature Extraction using an Unsupervised Neural Network

Nathan Intrator
Div. of Applied Mathematics, and Center for Neural Science
Brown University
Providence, RI 02912

Abstract

A novel unsupervised neural network for dimensionality reduction which seeks directions emphasizing distinguishing features in the data is presented. A statistical framework for the parameter estimation problem associated with this neural network is given and its connection to exploratory projection pursuit methods is established. The network is shown to minimize a loss function (projection index) over a set of parameters, yielding an optimal decision rule under some norm. A specific projection index that favors directions possessing multimodality is presented. This leads to a similar form to the synaptic modification equations governing learning in Bienenstock, Cooper, and Munro (BCM) neurons (1982).

The importance of a dimensionality reduction principle based solely on distinguishing features, is demonstrated using a linguistically motivated phoneme recognition experiment, and compared with feature extraction using principal components and back-propagation network.

1 How to construct optimal unsupervised feature extraction

When a classification of high dimensional vectors is sought, the *curse of dimensionality* (Bellman, 1961) becomes the main factor affecting the classification performance. The curse of dimensionality problem is due to the inherent sparsity of high dimensional spaces, implying that the amount of training data needed to get reasonably low variance estimators is ridiculously high. One approach to the problem is to assume that important structure in the data actually lies in a much smaller dimensional space, and therefore try to reduce the dimensionality before attempting the classification.

Hence, the desired property of a dimensionality reduction/feature extraction method is to lose as little information as possible after the transformation from the high dimensional space to the low dimensional one. This motivation underlies methods such as principal components (PC), mutual information maximization (Linsker, 1986), and self supervised form of back-propagation.

At a first glance, it seems that a supervised feature extraction method will always be superior to an unsupervised one, because if one has more information about the problem, it is natural to suppose that finding the solution is easier. However, unsupervised methods use a **local** measure to optimally estimate single dimensional functions of projections instead of functions of the full dimensionality of the space, and therefore tend to be less sensitive to the curse of dimensionality problem (Huber, 1985).

One way to reduce the curse of dimensionality is to look for lower dimensional structures (features) by using a localized and smooth objective function that directly measures the importance of the extracted features.

A useful class of features to explore is defined by some linear projections of the high dimensional data. This class is used in projection pursuit methods (PP) originally introduced by Kruskal (1969, 1972), Switzer (1970, 1971), and later implemented by Friedman and Tukey (1974). These methods are reviewed in Huber (1985).

It is still difficult to characterize what interesting projections are, although, it is easy to point at projections that are uninteresting. To motivate this discussion, consider the following example in which two data clusters lie in a two dimensional space. If we are inter-

ested in reducing the dimensionality of the data, and still retaining an indication on the structure, it is best to project the data onto the x axis, even though the variance of the projection to the y axis is larger.

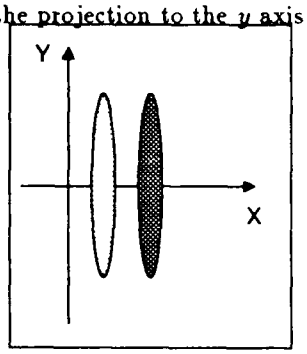

Figure 1: In this dimensionality reduction problem the interesting direction is not the one that maximizes the variance: Two data clusters which can be separated by projecting to the x axis, can not be separated by projecting to the y axis, although the variance in the y axis is larger.

Notice that in the above example, the projection onto the x axis will give a two hump distribution, while the projection onto the y axis will give a normal distribution. It turns out that this is not a coincidence. A statement that has recently been made precise by Diaconis and Freedman (1984) says that for most high-dimensional clouds, most low-dimensional projections are approximately normal. This finding suggests that the important information in the data is conveyed in those directions whose single dimensional projected distribution is far from Gaussian. Friedman (1987) argues that the most computationally attractive measures for deviation from normality (projection indices) are based on polynomial moments. For example, principal components extraction uses a projection index which is based on polynomials of the second moment of the projections (maximizing the projected variance). In some special cases where the data is known in advance to be bi-modal, it is relatively straightforward to define a good projection index (Hinton & Nowlan, 1990).

Despite their computational attractiveness, projection indices based on polynomial moments are not directly applicable, since they very heavily emphasize departure from normality in the tails of the distribution (Huber, 1985). Friedman (1987) addresses this issue by introducing a nonlinear transformation that squashes the projected data from R to $[-1, 1]$ using a normal distribution function. We address the problem by applying a sigmoidal squashing function to the projections, and then applying an objective function based on polynomial moments.

2 Feature Extraction using ANN

In this section, the intuitive idea presented above is used to form a statistically plausible objective function whose minimization will find those projections having a single dimensional projected distribution that is far from Gaussian.

We first informally describe the statistical formulation that leads to this objective function (the mathematical details are left to the appendix). Based on statistical decision theory, a neuron is considered as capable of making decisions. The most intuitive decision for a neuron is whether to fire or not for a given input and vector of synaptic weights. To aid the neuron in making the decision, a loss function is attached to each decision, namely a function that measures the loss from making each decision. The neurons task is then to choose the decision that minimizes the loss. Since the loss function depends on the synaptic weights vector in addition to the input vector, it is natural to seek a synaptic weight vector that will minimize the sum of the losses associated with every input, or more precisely, the average loss (also called the risk). The search for such vector, which yields an optimal synaptic weight vector under this formulation, can be viewed as learning or parameter estimation. In those cases where the risk is a smooth function its minimization can be done using gradient descent.

The ideas presented so far make no specific assumptions regarding the loss function, and it is clear that different loss functions will yield different learning procedures. For example, if the loss function is related to the inverse of the projection variance (including some normalization) then minimizing the risk will yield directions that maximize the variance of the projections, i.e. will find the principal components.

Before presenting our version of the loss function, let us review some necessary notations and assumptions. Consider a neuron with input vector $x = (x_1, \ldots, x_N)$, synaptic weights vector $m = (m_1, \ldots, m_N)$, both in R^N, and activity (in the linear region) $c = x \cdot m$. Define the threshold $\Theta_m = E[(x \cdot m)^2]$, and the functions $\hat{\phi}(c, \Theta_m) = c^2 - \frac{2}{3}c\Theta_m$, $\phi(c, \Theta_m) = c^2 - \frac{4}{3}c\Theta_m$. The ϕ function have been suggested as a biologically plausible synaptic modification function to explain visual cortical plasticity (Bienenstock, Cooper and Munro, 1982). The main features of BCM theory will be discussed below. Θ_m is a dynamic threshold which will be shown later to have an affect on the sign of the synaptic modification. The input x, which is a stochas-

tic process, is assumed to be of Type II φ mixing[1], bounded, and piecewise constant. These assumptions are plausible, since they represent the closest continuous approximation to the usual training algorithms, in which training patterns are presented at random. The φ mixing property allows for some time dependency in the presentation of the training patterns. The assumption are needed for the approximation of the resulting deterministic gradient descent by a stochastic one (Intrator, 1990b). For this reason we use a *learning rate* μ that has to decay in time so that this approximation is valid. Note that at this point c represents the *linear projection of x onto m, and we seek an optimal* projection in some sense.

Our projection index is aimed at finding directions for which the projected distribution is far from Gaussian, more specifically, we are interested in finding clusters in a high dimensional data. Since high dimensional clusters have a multimodal projected distribution, our aim is to find a projection index (loss function) that emphasizes multimodality. For computational efficiency, we would like to base the projection index on polynomial moments of low degree. Using second degree polynomials, one can get measures of the mean and variance of the distribution, which do not give information on multimodality, therefore, higher order polynomials are necessary. Furthermore, the projection index should exhibit the fact that bimodal distribution is already interesting, and any additional mode should make the distribution even more interesting.

With this in mind, consider the following family of loss functions which depend on the synaptic weight vector and on the input x (the derivation based on decision theory appears in the appendix).

$$\begin{aligned} L_m(x) &= -\mu \int_{\Theta_m}^{(x\cdot m)} \hat{\phi}(s,\Theta_m)ds \\ &= -\frac{\mu}{3}\{(x\cdot m)^3 - E[(x\cdot m)^2](x\cdot m)^2\} \end{aligned}$$

The motivation for this loss function can be seen in the following graph, which represents the ϕ function and the associated loss function $L_m(x)$. For simplicity the loss for a fixed threshold Θ_m and synaptic vector m can be written as $L_m(c) = -\frac{\mu}{3}c^2(c-\Theta_m)$, where $c = (x\cdot m)$.

[1]The φ mixing property specifies the dependency of the future of the process on its past.

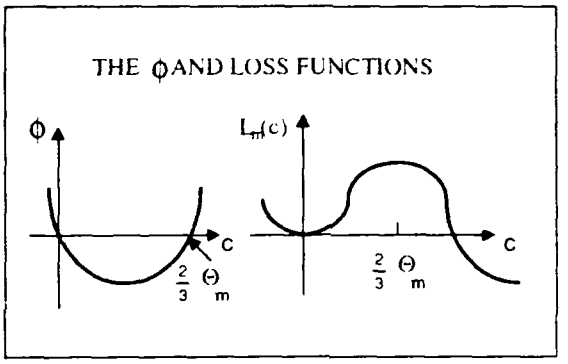

Figure 2: The function ϕ and the loss functions for a fixed m and Θ_m.

The graph of the loss function shows that for any fixed m and Θ_m, the loss is small for a given input x, when either $c = x\cdot m$ is close to zero, or when $x\cdot m$ is larger than $\frac{4}{3}\Theta_m$. Moreover, the loss function remains negative for $(x\cdot m) > \frac{4}{3}\Theta_m$, therefore any kind of distribution at the right hand side of $\frac{4}{3}\Theta_m$ is possible, and the preferred ones are those which are concentrated further from $\frac{4}{3}\Theta_m$.

It remains to show why it is not possible that a minimizer of the average loss will be such that all the mass of the distribution will be concentrated in one of the regions. Roughly speaking, this can not happen because the threshold Θ_m is dynamic and depends on the projections in a nonlinear way, namely, $\Theta_m = E(x\cdot m)^2$. This implies that Θ_m will always move itself to a position such that the distribution will never be concentrated at only one of its sides. This yield that the part of the distribution for $c < \frac{4}{3}\Theta_m$ has high loss, making those distributions in which the distribution for $c < \frac{4}{3}\Theta_m$ has its mode at zero, more plausible.

The fact that the distribution has part of its mass on both sides of $\frac{4}{3}\Theta_m$ makes it already a plausible projection index that seeks multi- modalities. However, this projection index will be more general, if in addition, the loss will be insensitive to outliers, if we allow any projected distribution to be shifted so that the part of the distribution that satisfies $c < \frac{4}{3}\Theta_m$ will have its mode at zero. These points will be discussed below.

The risk (expected value of the loss) is given by:

$$\begin{aligned} R_m &= -\tfrac{\mu}{3}E\{(x\cdot m)^3 - E[(x\cdot m)^2](x\cdot m)^2\} \\ &= -\tfrac{\mu}{3}\{E[(x\cdot m)^5] - E^2[(x\cdot m)^2]\}. \end{aligned}$$

Since the risk is continuously differentiable, its minimization can be achieved via a gradient descent

method with respect to m, namely:

$$\frac{dm_i}{dt} = -\frac{\partial}{\partial m_i} R_m = \mu \{E[(x \cdot m)^2 x_i] - \frac{4}{3} E[(x \cdot m)^2] E[(x \cdot m) x_i]\} = \mu E[\phi(x \cdot m, \Theta_m) x_i].$$

The resulting differential equations suggest a modified version of the law governing synaptic weight modification in the BCM theory for learning and memory (Bienenstock, Cooper and Munro, 1982). This theory was presented to account for various experimental results in visual cortical plasticity. According to this theory, the synaptic efficacy of active inputs increases when the postsynaptic target is concurrently depolarized beyond a *modification threshold*, Θ_m. However, when the level of postsynaptic activity falls below Θ_m, then the strength of active synapses decreases. An important feature of this theory is that the value of the modification threshold is not fixed, but instead varies as a nonlinear function of the average output of the postsynaptic neuron. This feature provides the stability properties of the model, for positive or mean positive inputs, and is necessary in order to explain, for example, why the low level of postsynaptic activity caused by binocular deprivation does not drive the strengths of all cortical synapses to zero. Mean field theory for a network based on these neurons is presented in (Scofield and Cooper, 1985; Cooper and Scofield, 1988), statistical analysis is given in Intrator (1990c) computer simulations and biological relevance are discussed in (Soul et al., 1986; Bear et al., 1987; Cooper et al., 1987; Bear et al., 1988; Clothioux, 1990).

Up to this point we have presented an unsupervised (exploratory) method for feature extraction that seeks projections in which the single dimensional distribution is multi-modal, namely we have presented an exploratory projection pursuit method. This method uses polynomial moments as a projection index and therefore suffers from over-sensitivity to outliers (Freidman, 1987). We address this problem by considering a nonlinear neuron in which the neuron's activity is defined to be $c = \sigma(x \cdot m)$, where σ usually represents a smooth sigmoidal function. A more general definition that would allow symmetry breaking of the projected distributions, will provide solution to the second problem raised above, and is still consistent with the statistical formulation is $c = \sigma(x \cdot m - \alpha)$, for an arbitrary threshold α which can be found by using gradient descent as well. For the nonlinear neuron Θ_m is defined to be $\Theta_m = E[\sigma^2(x \cdot m)]$. The loss function is given by:

$$L_m(x) = -\mu \int_{\Theta_m}^{\sigma(x \cdot m)} \hat{\phi}(s, \Theta_m) ds = -\frac{\mu}{3} \{\sigma^3(x \cdot m) - E[\sigma^2(x \cdot m)]\sigma^2(x \cdot m)\}$$

The gradient of the risk becomes:

$$-\nabla_m R_m = \mu \{E[\sigma^2(x \cdot m)\sigma' x] - \frac{4}{3} E[\sigma^2(x \cdot m)] E[\sigma(x \cdot m)\sigma' x]\} = \mu E[\phi\Big(\sigma(x \cdot m), \Theta_m\Big)\sigma' x],$$

where σ' represents the derivative of σ at the point $(x \cdot m)$. Note that the multiplication by σ' reduces sensitivity to outliers of the differential equation since for outliers σ' is close to zero.

Based on this formulation, a network of Q identical nodes, which receive the same input and inhibit each other, may be constructed in order to extract several features at once. A similar network has been studied by Scofield and Cooper (1985). The activity of neuron k in the network is defined as $c_k = x \cdot m_k$, where m_k is the synaptic weight vector of neuron k. The *inhibited* activity and threshold of the k'th neuron are given by

$$\tilde{c}_k = c_k - \eta \sum_{j \neq k} c_j, \qquad \tilde{\Theta}_m^k = E[\tilde{c}_k^2].$$

Schematic structure of the network is given in Figure 3.

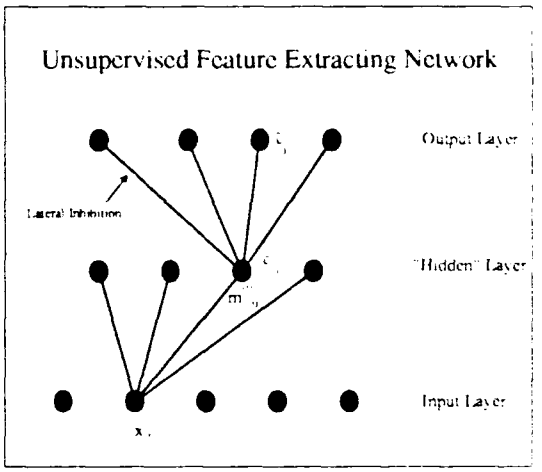

Figure 3: The activity of a nonlinear neuron j is given by $c_j = \sigma(x \cdot m_j)$, the inhibited activity is given by $\tilde{c}_j = c_j - \eta \sum_{k \neq j} c_k$.

We omit the derivation of the synaptic modification equations which is similar to the one for a single neuron, and present only the resulting modification equations for a synaptic vector m_k in a lateral inhibition network of nonlinear neurons:

$$\dot{m}_k = -\mu E\{\phi(\tilde{c}_k, \tilde{\Theta}_m^k)\Big(\sigma'(x \cdot m_k)$$

$$-\eta \sum_{j \neq k} \sigma'(x \cdot m_j)\Big) x\}$$

The full derivation can be found in Intrator (1990a). The lateral inhibition network performs a direct search of k-dimensional projections together, which may find a richer structure that a stepwise approach may miss, e.g. see example 14.1 Huber (1985).

3 Comparison with other feature extraction methods

The problem of feature extraction for classification is in some sense easier than that of feature extraction for density or function estimation. This is because the only interesting features in such case are those that distinguish *between* a finite set of classes. The common features, namely those features that do not help in making the distinction between classes are uninteresting, even though they may be very important for data compression, e.g. the self supervised back-propagation network in which the number of hidden units is smaller than the number of input and output units (Elman & Zipser, 1989). The network presented in the previous sections has been shown to seek multimodality in the projected distributions, which translates to clusters in the original space, and therefore to find those directions that make a distinction between different sets in the training data.

In this section we explore the differences in classification performance between a network that performs dimensionality reduction (before the classification) based upon distinguishing features, and a network that performs dimensionality reduction based upon minimization of misclassification error. The performance of the different methods will be compared on a specific classification task: a phoneme classification experiment whose linguistic motivation is described below.

We looked at the six stop consonants [p,k,t,b,g,d] which have been a subject of recent research in evaluating neural networks for phoneme recognition (see review in Lippmann, 1989). These stops posses several common features, but only two distinguishing phonetic features, place of articulation and voicing (table 1) (see Blumstein & Lieberman for a review and related references on phonetic feature theory).

	Place of Articulation		
	Velar	Alveolar	Labial
Voiced	[g]	[d]	[b]
Unvoiced	[k]	[t]	[p]

Table 1: The two distinguishing phonetic features between the six stop consonants.

The Linguistic information in the table suggests the following experiment: A network is to be trained to reduce dimensionality from the unvoiced stops [p,k,t]. In order to reduce variability in the data, only a single speaker and a single vowel context is used. Therefore, the only distinguishing features in the training data are associated with place of articulation, since the features that are speaker dependent, voicing dependent, or context dependent belong to the set of common features in the training data. A dimensionality reduction method that concentrates mainly on distinguishing features should find only the features associated with place of articulation, and therefore become insensitive to voicing dependent and speaker dependent features, which are the common features in the training data. This can easily be tested by evaluating the performance on place of articulation classification of voiced stops and data from other speakers.

For comparison, we have attempted to extract features using three methods: principal components, back-propagation, and the above unsupervised network, all trained and tested on the same data. In back-propagation, the only supervised method, the place of

articulation phonetic feature was used as a supervisor.

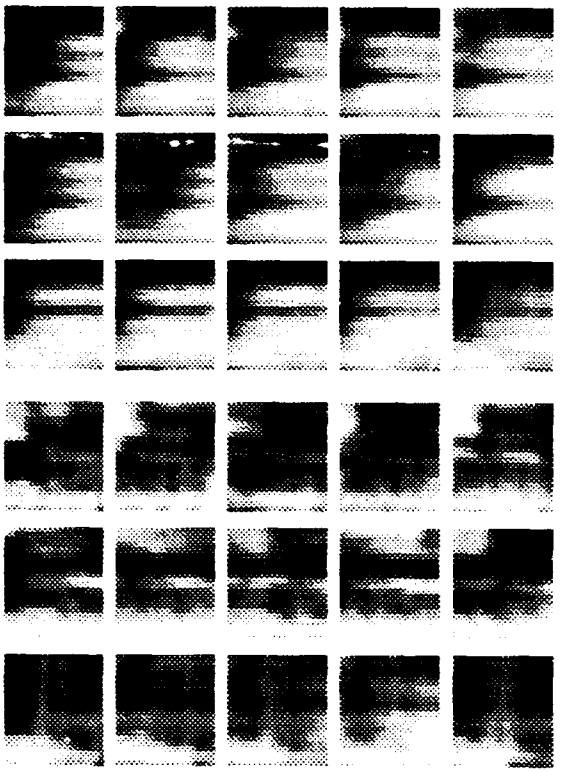

Figure 4: The six stop consonants followed by the vowel [a] for male speaker BSS. Their order from bottom to top is [pa] [ka] [ta] [ba] [ga] [da]. Each token is represented by a 20 consecutive time windows of 32msec with 30msec overlap. In each time frame a set of 22 energy levels in Zwicker critical band filters are computed. Notice the significant difference between the voiced and the unvoiced images.

The speech data consists of 20 consecutive time windows of 32msec with 30mSec overlap, aligned to the beginning of the burst. In each time window, a set of 22 energy levels is computed. These energy levels correspond to Zwicker critical band filters (Zwicker, 1961).

The consonant-vowel (CV) pairs were pronounced in isolation by native American speakers (two male BSS and LTN, and one female JES.) Five tokens of each of the CV pairs used for training are presented in Figure 4. Additional details on biological motivation for the preprocessing, and linguistic motivation related to child language acquisition can be found in Seebach (1990), and Seebach and Intrator (1990).

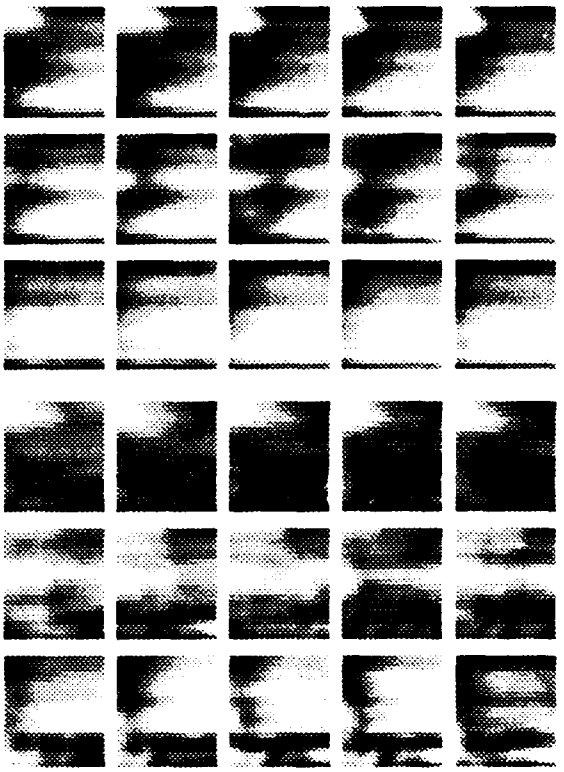

Figure 5: The six stop consonants followed by the vowel [a] for female speaker JES. Their order from bottom to top is [pa] [ka] [ta] [ba] [ga] [da]. Pre-processing is the same as above. Notice that the same burst that appear in [ta] is clear in the [da] as well.

Figure 5 presents five tokens of each of the CV pairs pronounced by the female speaker JES. The classification results obtained using BCM network and principal components methods, were better on this speaker, than on those obtained when testing the performance on the speaker that was used in the training. This is due to the very 'clean' sound that corresponds closely to the acoustic features that are known (Blumstein & Lieberman, 1984) to exist in these sounds. For example, this was the only speaker out of several that we tested, in which the high frequency burst (top left corner) is clear for the voiced stop as it is clear for the unvoiced stops.

The unsupervised feature extraction/classification method is presented in Figure 6. Similar approach using the RCE and back-propagation network have been carried out by several researchers (Rimey et al., 1986; Reilly et al., 1987, 1988; Zemani et al., 1989), and using the unsupervised charge clustering network by Scofield (1988)

Five features/directions were extracted from the 440 dimensional preprocessed speech vectors. These features were the activation of five neurons in the unsupervised network, the five principal components in the PC method, and the five hidden unit activations in back-propagation. The extracted features were used to train a k-NN classifier (with $k = 3$) to classify place of articulation. Although the three dimensionality reduction methods were trained only with the unvoiced tokens of a single speaker, the five dimensional k-NN classifier was trained on voiced and unvoiced data from the other speakers as well.

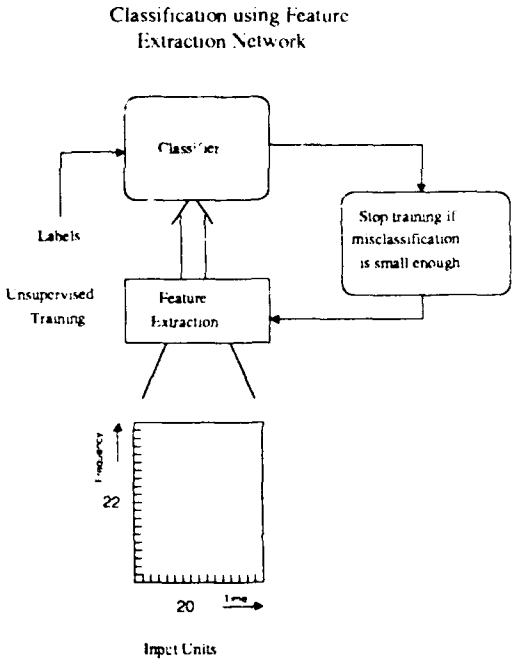

Figure 6: Low dimensional k-NN classifier is trained on the features extracted from the high dimensional data. Training of the feature extraction network stops, when misclassification rate drops below a predetermined threshold on either the same training data (cross validatory test) or on a different testing data.

The classification results are summarized in table 2. Several observations can be made from the results; First, the principal components dimensionality reduction is clearly not sufficient in discovering structure for this kind of data, suggesting that the structure is highly non linear. Second, the back-propagation network is doing well in finding structure useful for classification of the trained data, but this structure does not concentrates on distinctive features solely, it also contains speaker dependent and voicing dependent features, and therefore has degraded classification performance when tested on voiced data, or data from other speakers. This can also be viewed as a generalization problem, in which case one can say that the network is overfitting to the training data. Third, classification results using the BCM network for dimensionality reduction suggest that for this specific task, structure that is less sensitive to voicing features can be extracted, even though the network was trained on the unvoiced data only and voicing has significant effects on the speech signal itself.

Place of Articulation Classification			
	P-C	B-P	BCM
BSS /p,k,t/	66.0	100.0	98.6
BSS /b,g,d/	57.4	73.3	94.0
LTN /p,k,t/	60.0	95.8	98.9
LTN /b,g,d/	46.6	66.7	90.0
JES (Both)	70.6	83.7	99.4

Table 2: Percentage of correct classification of place of articulation in voiced and unvoiced stops using principal components, back-propagation, and the BCM network. Training for dimensionality reduction was done on unvoiced stops of male speaker BSS in all three experiments. LTN is a male speaker aswell. The result in the last column represents testing on both the voiced and unvoiced stops of a female speaker (JES). The results represent an average result of several trials, which differ only in the initial conditions of the networks.

4 Discussion

It has been shown that the BCM neuron is capable of effectively discovering nonlinear structures in high dimensional spaces. When compared with other projection indices, the highlights of the presented method are *i*) the projection index concentrates on directions where the separability property as well as the non-normality of the data is large, thus giving rise to bet-

ter classification properties; *ii*) the degree of correlation between the directions (features) extracted by the network can be regulated via the global inhibition, allowing some tuning of the network to different types of data for optimal results; *iii*) the pursuit is done on all the directions at once thus leading to the capability of finding more interesting structures than methods that find only one projection direction at a time.

Regarding the speech experiment, the network and its training paradigm present a different approach to speaker independent speech recognition. In this approach the speaker variability problem is addressed by training a network that concentrates mainly on the distinguishing features, on a single speaker, as opposed to training a network that concentrates on both the distinguishing and common features, on multi-speaker data.

Acknowledgements

I wish to thank Leon N Cooper for suggesting the problem and for providing many helpful hints and insights. Geoff Hinton made invaluable comments that made this manuscript much more readable. The application of BCM to speech is discussed in more detail in Seebach (1990) and in a forthcoming article (Seebach and Intrator, 1990). The back-propagation experiments were done by Charles M. Bachmann.

Research was supported by the Office of Naval Research, the National Science Foundation, and the Army Research Office.

References

Bellman, R. E. (1961) Adaptive Control Processes. Princton, NJ, Princton University Press.

Bienenstock, E. L. (1980) A theory of the development of neuronal selectivity. Doctoral dissertation, Brown University, Providence, RI

Bienenstock, E. L., L. N Cooper, and P.W. Munro (1982) Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *J.Neurosci.* 2:32-48

Bear, M. F., L. N Cooper, and F. F. Ebner (1987) A Physiological Basis for a Theory of Synapse Modification. *Science* 237:42-48

Bear, M. F., L. N Cooper, and F. F. Ebner (1988) Synaptic Modification Model of Learning and Memory. In *Encyclopedia of Neuroscience.*

Cooper, L. N and C. L. Scofield (1988) Mean-field theory of a neural network. *Proc. Natl. Acad. Sci. USA* 85:1973-1977

Devijver P. A., and J. Kittler (1982) Pattern Recognition: A Statistical Approach. Prentice Hall London

Diaconis, P, and D. Freedman (1984) Asymptotics of Graphical Projection Pursuit. *The Annals of Statistics*, 12 793-815.

Friedman, J. H. and J. W. Tukey (1974) A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.* C-23:881-889

Friedman, J. H. and W. Stuetzle (1981) Projection pursuit regression. *J. Amer. Statist. Assoc.* 76:817-823

Friedman, J. H., W. Stuetzle and A. Schroeder (1984) Projection pursuit density estimation. *J. Amer. Statist. Assoc.* 79:599-608

Friedman, J. H. (1987) Exploratory Projection Pursuit. *Journal of the American Statistical Association* 82-397:249-266

Hall, P. (1988) Estimating the Direction in which Data set is Most Interesting. *Probab. Theory Rel. Fields* 80, 51-78

Hall, P. (1989) On Polynomial-Based Projection Indices for Exploratory Projection Pursuit. *The Annals of Statistics*, 17, 589-605.

Hinton, G. E. and S. J. Nowlan (1990) *Neural Computation*, (in press).

Huber P. J. (1981) Projection Pursuit. Research Report PJH-6, Harvard University, Dept. of Statistics.

Huber P. J. (1985) Projection Pursuit. *The Annal. of Stat.* 13:435-475

Intrator N. (1990) A Neural Network For Feature Extraction. In D. S. Touretzky (ed.), *Advances in Neural Information Processing Systems 2.* San Mateo, CA: Morgan Kaufmann.

Intrator N. (1990) An Averaging Result for Random Differential Equations. Technical report CNS-54, Brown University.

Intrator N. (1990) Feature Extraction using an Exploratory Projection Pursuit Neural Network. Ph.D. Dissertation Brown University.

Jones, M. C. (1983) The Projection Pursuit Algorithm for Exploratory Data Analysis. Unpublished Ph.D. dissertation, University of Bath, School of Mathematics.

Kruskal, J. B. (1969) Toward a practical method which helps uncover the structure of the set of multivariate observations by finding the linear transformation which optimizes a new "index of condensation". In *Statistical Computation.* (R.C. Milton and J. A. Nelder,

eds.)

Kruskal, J. B. (1972) Linear Transformation of multivariate data to reveal clustering. In *Multidimensional Scaling: Theory and Application in the Behavioral Sciences, I, Theory.* Seminar Press, New York and London.

Linsker R. (1988) Self-Organization in a Perceptual Network. *IEEE. Computer* March 88:105-117.

Lippmann, R. P. (1989) Review of Neural Networks for Speech Recognition. *Neural Computation* 1, 1-38.

von der Malsburg, C. (1973) Self-organization of orientation sensitivity cells in the striate cortex. *Kybernetik* 14:85-100

Reilly, D. L., C.L. Scofield, C. Elbaum and L. N Cooper (1987) Learning system architectures composed of multiple learning modules *Proc. First International Conference on Neural Networks.*

Reilly, D. L., C.L. Scofield, L. N Cooper and C. Elbaum (1988) GENSEP: a multiple neural network with modifiable network topology. *INNS Conference on Neural Networks.*

Rimey, R., P. Gouin, C.L. Scofield and D. L. Reilly (1986) Real-time 3-D object classification using a learning system. *SPIE, Intelligent Robots and Computer Vision*, October 1986.

Saul, A. and E. E. Clothiaux, 1986) Modeling and Simulation II: Simulation of a Model for Development of Visual Cortical specificity. *J. of Electrophysiological Techniques*, 13:279-306

Scofield, C. L. and L. N Cooper (1985) Development and properties of neural networks. *Contemp. Phys.* 26:125-145

Scofield, C. L. (1988) Unsupervised learning in the N dimensional Coulomb net. *Abstracts in the first annual international neural network soc. meeting.* Vol. 1 Supl. 1, 1988 p.129.

D. P. Morgan, C. L. Scofield. T.M. Lorenzo, E.C Real and D.P Loconto (1990) A key word spotter which incorporates neural network for secondary processing. *Proc. ICCASSP*, Alberquerque New Mexico 1990 pp.113-116.

Seebach, B. S. (1990) Evidence for the Development of Phonetic Property Detectors in a Neural Net without Innate Knowledge of Linguistic Structure. Ph.D. Dissertation Brown University.

Seebach, B. S. and N. Intrator (1990) To appear.

Switzer, P. (1970) Numerical classification. IN *Geostatistics.* Plenum, New York.

Switzer, P. and R. M. Wright (1971) Numerical classification applied to certain Jamaican eocene nummulitids. *Math. Geol.* 3:297-311

Zemani P. D., D. P. Morgan, D. L. Reilly, C. L. Scofield et al. (1989) Experiments in discrete utterance recognition using neural network. *Proc. of the Poston ASSP mini-conference, Weston Massachusetts* May, 1989

Zwicker E. (1961) Subdivision of the audible frequency range into critical bands (Frequenzgruppen) *Journal of the Acoustical Society of America* 33:248

Mathematical Appendix

In this section we develop the statistical formulation that yields the loss function presented in section 2.

Let $(\Omega, \mathcal{F}_\Omega, P)$ be a probability space on the space of inputs Ω with probability law P. Let $\mathcal{A} = \{0, 1\}$ be a decision space, in the case of a single neuron a zero decision means that the neuron does not fire. Let m be a vector of parameters such as the one described above, and assume that it lies in a compact space B^M. This parameter space defines a family of loss functions, $\{L_m\}_{m \in B^M}$, $L_m : \Omega \times \mathcal{A} \mapsto R$. Let $\mathcal{D}$ be the space of all decision rules. The empirical risk (average loss) $R_m : \mathcal{D} \mapsto R$, is given by:

$$R_m(\delta) = \sum_{i=1}^{n} P(x^{(i)}) L_m(x^{(i)}, \delta(x^{(i)}))$$

For a fixed m, the optimal decision δ_m is chosen so that:

$$R_m(\delta_m) = \min_{\delta \in \mathcal{D}} \{R_m(\delta)\}$$

Since this minimization takes place over a finite set, the minimizer exists. In particular, for a given $x^{(i)}$ the decision $\delta_m(x^{(i)})$ is chosen so that

$$L_m(x^{(i)}, \delta_m(x^{(i)})) \leq L_m(x^{(i)}, 1 - \delta_m(x^{(i)})).$$

At this point $R_m(\delta_m)$ is a risk function that depends only on the vector of parameters m, and assuming R_m is bounded, it is natural to seek a parameter $\bar{m}$ that minimizes R_m, namely,

$$R_{\bar{m}}(\delta_{\bar{m}}) = \min_{m \in B^M} \{R_m(\delta_m)\}.$$

The minimum with respect to m exists since B^M is compact, and R_m is bounded. When m represents a vector in R^N, R_m can be viewed as a projection index.

Based on the above, let m, the synaptic weight vector, be the parameter to be esimated, and consider the following family of loss functions. The loss functions depend on the cell's decision whether to fire or not, and

they represent the intuitive idea that the neuron will fire when its activity is greater than some threshold, and will not otherwise. We denote the firing of the neuron by $a = 1$. Define $K = -\mu \int_0^{\frac{2}{3}\Theta_m} \hat{\phi}(s, \Theta_m) ds$. The loss function for a decision to fire is given by:

$$L_m(x, 1) = \begin{cases} -\mu \int_{\Theta_m}^{(x \cdot m)} \hat{\phi}(s, \Theta_m) ds, & (x \cdot m) \geq \Theta_m \\ K - \mu \int_{\Theta_m}^{(x \cdot m)} \hat{\phi}(s, \Theta_m) ds, & (x \cdot m) < \Theta_m, \end{cases}$$

and for the decision not to fire by:

$$L_m(x, 0) = \begin{cases} -\mu \int_{\Theta_m}^{(x \cdot m)} \hat{\phi}(s, \Theta_m) ds & (x \cdot m) \leq \Theta_m \\ K - \mu \int_{\Theta_m}^{(x \cdot m)} \hat{\phi}(s, \Theta_m) ds, & (x \cdot m) > \Theta_m. \end{cases}$$

It follows from the definition of L_m and from the definition of δ_m that

$$\begin{aligned} L_m(x, \delta_m) &= -\mu \int_{\Theta_m}^{(x \cdot m)} \hat{\phi}(s, \Theta_m) ds \\ &= -\frac{\mu}{3} \{(x \cdot m)^3 - E_i(x \cdot m)^2{}_i (x \cdot m)^2\}. \end{aligned}$$

We can write $L_m(x)$ instead of $L_m(x, \delta_m)$ when there is no confusion.

The risk is given by:

$$R_\theta(\delta_\theta) = -\frac{\mu}{3} \{E_i(x \cdot m)^3{}_i - E^2{}_i(x \cdot m)^2{}_i\}.$$

Since the risk is continuously differentiable, its minimization can be done via the gradient descent method with respect to m, namely:

$$\frac{\partial m_i}{\partial t} = -\frac{\partial}{\partial m_i} R_\theta(\delta_\theta) = \mu \, E_i \phi(x \cdot m, \Theta_m) x_i{}_i.$$

Notice that the resulting equation represents an averaged deterministic equation of the stochastic BCM modification equations. It turns out that under suitable conditions on the mixing of the input x and the global function μ, this equation is a good approximation of its stochastic version (Intrator, 1990b), namely:

$$\frac{\partial m_i}{\partial t} = \mu \, \phi(x \cdot m, \Theta_m) x_i.$$